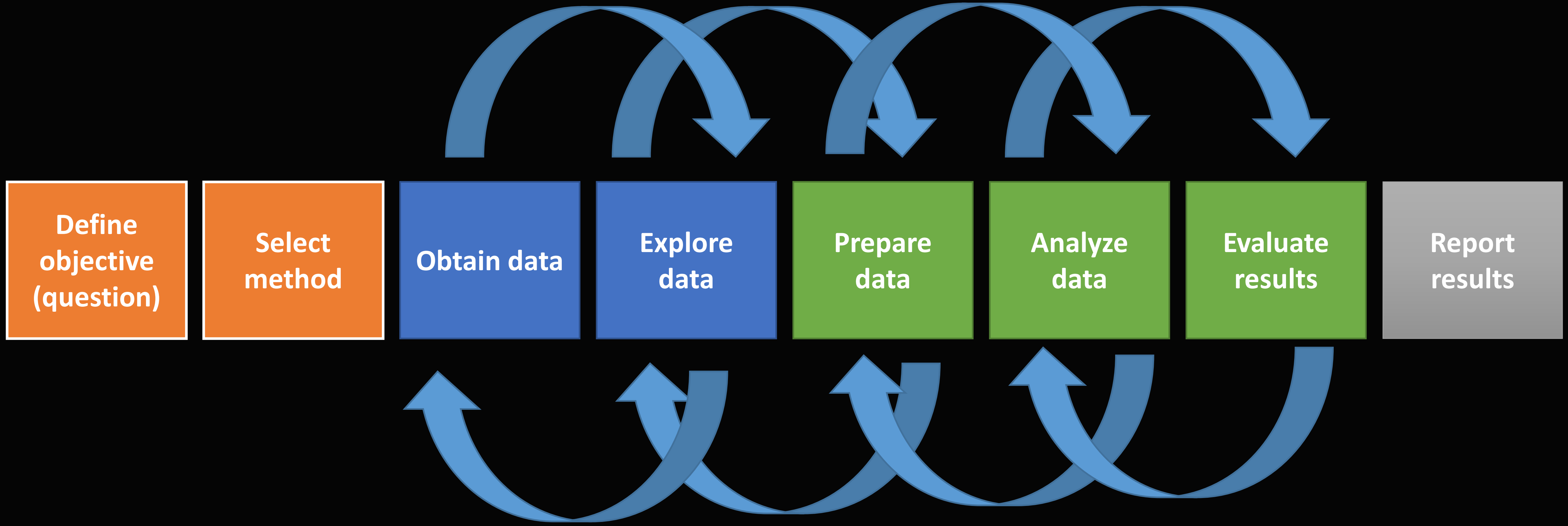
STATISTICAL RESEARCH DESIGN

Zack Kertcher

Statistics for Management
Fall 2016

Statistical research process



Define objective (question)

Select method

Obtain data

Explore data

Prepare data

Analyze data

Evaluate results

Report results

Plan for today

1. Research question
2. Research methodology
3. Reporting results
4. Review of team project/final project
5. *Data prep (messy data)

THE RESEARCH QUESTION

Examples of research questions

1. Clear, empirically driven question

✓ Has to make sense and be clear
(Specify what you are trying to describe and predict)

✓ Based on empirical evidence, data that can be verified
(Not based on speculation, moral judgement, subjective pref.)

Examples

Can ACME Car Dealership be more profitable this year?

Can transactions in ACME Car Dealership Midwest this year, be more profitable compared to last year?

What factors effect the profitability of transactions at ACME Car Dealership Midwest?

Examples

Can ACME Car Dealership we be more profitable this year?

Can transactions in ACME Car Dealership Midwest this year, be more profitable compared to last year?

What factors effect the profitability of transactions at ACME Car Dealership Midwest?

2. Use hypotheses to guide the research

✓ A formal statement about a relationship between variables

✓ Guides the research by offering **tentative predictions**
(about a relationship between variables)

✓ Typically relates to an **independent** (x) and **dependent variable** (y)
(A change in x will negatively effect y)

✓ Can be empirically tested
(Using data and statistical inferential methods)

Examples

Low-performing salespeople at ACME Midwest sell less.

Comparing the salesforce in ACME Midwest, some salespeople are likely to sell more than others.

When looking at salespeople at ACME Midwest, experience and age are related.

The profitability from sales at ACME Midwest is higher for salespeople who successfully completed corporate training.

Examples

Low-performing salespeople at ACME Midwest sell less.

Comparing the salesforce in ACME Midwest, some salespeople are likely to sell more than others.

When looking at salespeople at ACME Midwest, experience and age are related.

The profitability from sales at ACME Midwest is higher for salespeople who successfully completed corporate training.

- ✓ Select a research question pertaining to a business scenario
- ✓ Formulate hypotheses that address this research question

RESEARCH METHODOLOGY

What to consider when selecting a research method?

1. Does it fully support my research question?

If the question is about a causal relationship, establish that:

a. The cause preceded the effect in time
b. There is an empirical relationship been cause and effect
c. Their relations cannot be explained by any other variable

Conduct an experiment (random assignment, control group)

If the question is about the effect of a variable(s) on an outcome, consider:

a. Is the outcome time-dependent?
    - No: A cross-sectional survey, sample/complete data
    - Yes: A panel survey, time-series data

Examples

Does corporate training cause increased profitability?

What factors have a positive effect on profitability?

Training was provided every month in the past 10 months. What is the predicted effect profitability?

Examples

Does corporate training cause increased profitability?

Experiment

What factors have a positive effect on profitability?

Survey/Cross-sectional data

Training was provided every month in the past 10 months. What is the predicted profitability increase in the next 2 months?

Time-series data

- ✓ Select two research methods to answer the research question you have developed above.

- ✓ Which one do you think would be the best for answering your research question?

2. What are the units of analysis?

- Classify the variable you intend to analyze as IV, DV, or CV.

- Dependent variable (DV). What you are trying to explain. Typically just one.
- Independent variable (IV). The variable that influences/causes (directly or indirectly) the dependent variable. Typically multiple, but often the focus is on one.

IV → DV

- Control variable (CV). "Generic" variables that often effect DV. You want to "control" (hold these constant) when examining the relationship between the IV(s) and DV. Often control variables are demographic in nature, such as gender, age, location.

2. What are the units of analysis? (cont.)

✔ Determine the unit of analysis
Individuals vs. groups (employee vs. team vs. department vs. city vs. region)

✔ Consider how each one is measured?
Character, factor, numeric (ordinal/discrete), numeric (continuous)

✔ Establish additional research design decisions
Including duration of data, sample/subject size, data structure

Example

- ✓ Research question: What factors effect profitability?
- ✓ Identify the IV, DV and control variables
- ✓ What units of measurement are appropriate for each?

profitability, age, sales_total, region, gender, months_with_company, corporate_training, months_since_training

Example

- ✓ Research question: What factors effect profitability?
- ✓ Identify the IV, DV and control variables.
- ✓ What units of measurement are appropriate for each?

profitability (DV), age (CV), sales_total (IV), region (CV/IV), gender (CV), months_with_company (CV/IV), corporate_training (IV), months_since_training (IV)

Example

- ✓ Research question: What factors effect profitability?
- ✓ Identify the IV, DV and control variables.
- ✓ What units of measurement are appropriate for each?

profitability (DV-numeric: 0-1), age (CV-numeric: 18-70),
sales_total (IV-numeric: 0-1000), region (CV/IV-factor: East/Midwest/South/West), gender (CV-factor: M/F),
months_with_company (CV/IV-numeric: 0-600),
corporate_training (IV-factor: Y/N),
months_since_training (IV-numeric: 0-300)

- ✓ Determine the units of analysis in your research

- ✓ Include DV, IV and control variables, as well as their measurement

REPORT RESULTS

What to report?

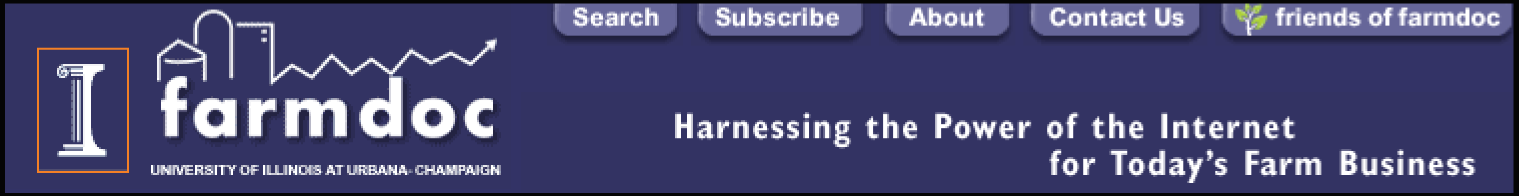
- ✓ Include a high-level summary
  (This is the abstract/executive summary–briefly, review your project)

- ✓ Explain data and analytical procedures
  (Acknowledge limitations)

- ✓ Show, interpret and explain results

- ✓ Conclude
  (What's the take home message?)

Example


Search
Subscribe
About
Contact Us
friends of farmdoc
farmdoc
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
Harnessing the Power of the Internet
for Today's Farm Business

http://farmdoc.illinois.edu/podcasts/mobr/MOBR_11_02.mp3

How to report text+plots+code?

Copy/Paste

Code/Process

- ✓ In RStudio File -> New File -> RMarkdown…

- ✓ Load the baseball data (from the midterm exam)

- ✓ Answer question 6, 7 or 8 from the midterm exam, in a report format (Hint: You should have the answer. Just "plug" it into RMarkdown

- ✓ Click on Knit HTML
  (To Knit PDF or Word you will need to install several packages)

Additional resources

rmarkdown

http://www.markdowntutorial.com/

https://www.rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf

ReporteRs

http://davidgohel.github.io/ReporteRs/index.html

FINAL PROJECT

Logistics (and Q&A)

Team?

- ✓ Teams up to 6 students (4-5 seem to work best)
- ✓ You get to choose your own group
- ✓ We can help "match" you if needed

Grade?

- ✓ 25% or 33% (depending if you take the final)
- ✓ 20% presentation (11/19), 80% final report (12/10)

Data and analysis

- ✓ You find data (preferably original data)
- ✓ Formulate a research question, hypotheses
- ✓ Analyze the data and report finding

*DATA PREP (MESSY DATA)

Common problems

Problem
Messy column names
Multiple columns for a single variable
Missing data (blanks, NA)
Odd data values (age = -1)
Odd characters (price = $70,000)
Data/time formatting

Common problems

Problem	Solution
Messy column names	Change them
Multiple columns for a single variable	Reshape data using tidyr or reshape2. Every column=variable, row=observation.
Missing data (blanks, NA)	Convert blanks to NA. Decide what to do with NAs (e.g., drop them, but could be a problem if there are too many)
Odd data values (age = -1)	Explore meaning, and decide what to do (e.g., -1 likely a data entry error, or missing value).
Odd characters (price = $70,000)	Remove them using string matching, or other methods (e.g., gsub)
Data/time formatting	Use as.Date and related base functions and lubridate